SONHICA (Simple Optimized Non-HIerarchical Cluster Analysis): A New Tool for Analysis of Molecular Conformations

GIANPAOLO BRAVI, EMANUELA GANCIA, ANDREA ZALIANI, MONICA PEGNA

Italfarmaco Research Centre, via Lavoratori 54, 20092 Cinisello Balsamo (Milano), Italy

Received 28 June 1996; revised 2 January 1997

ABSTRACT: We describe a new clustering program, SONHICA (Simple Optimized Non-HIerarchical Cluster Analysis), developed to analyze large data sets of molecular conformations. Unlike traditional clustering methods, SONHICA does not make use of an overall index, like a distance, to evaluate similarity between objects. Each descriptor variable is compared individually on the basis of a preset threshold value. This assures high control and sensitivity over the input variables. In addition, periodic and nonperiodic descriptors, such as dihedral angles and interatomic distances, can easily be used together. SONHICA generates clusters with the highest possible density and all pairs of objects within a cluster are similar. These features make SONHICA particularly suitable for the analysis of data sets which tend to form globular clusters. This method was applied to the analysis of a modified linear tetrapeptide, ITF1697, under investigation for its anti-ischemic properties, and a cyclic pentapeptide, BQ123, a potent antagonist of endothelin A. On the basis of the results presented here, SONHICA appears to be an interesting new tool in the field of the clustering methods applied to the analysis of molecular conformations. © 1997 by John Wiley & Sons, Inc. *J Comput Chem* **18:**1295–1311, 1997

Keywords: cluster analysis; SONHICA; conformational analysis; ITF1697; BQ123; molecular dynamics

Correspondence to: G. Bravi, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285-0540; e-mail: bravi-gianpaolo@lilly.com

CCC 0192-8651 / 97 / 101295-17

Introduction

Molecular dynamics, distance geometry, and stochastic and systematic search methods are widely used to investigate the conformational space of a molecule.[1] The choice of the most appropriate method depends strongly upon the nature of the chemical structures to be examined. For example, when studying molecules with a high degree of conformational freedom, like peptides, molecular dynamics simulations or stochastic methods are preferred to systematic ones. Although this may require the generation of a large number of molecular conformations, the use of modern equipment considerably reduces the otherwise extensive computational time. However, what is often difficult, tedious, and very time-consuming is the analysis of such data sets, and in particular the definition of possible conformational classes.

Chemometrics provides us with several tools which simplify the interpretation of large multivariate data sets. Cluster analysis[2–4] is widely used among unsupervised pattern recognition techniques, in particular when the objects correspond to molecular conformations and when the descriptor variables are their atomic coordinates, interatomic distances, or dihedral angles. Thus, cluster analysis permits the identification of ''natural'' groupings of data and the extraction of additional information about the relationships between different groups. It can therefore be used either as a method of exploratory data analysis or to derive classification models for structure–property relationship studies.

We describe here a new clustering program, called SONHICA (Simple Optimized Non-HIerarchical Cluster Analysis), which proved to be extremely effective in the analysis of conformational data sets.

The details of the algorithm are described and its properties are illustrated by means of two applications: the first on ITF1697,[5] a modified linear tetrapeptide under development as an antiischemic molecule; and the second on BQ123,[6] a potent cyclic pentapeptide antagonist of endothelin A.

Cluster Analysis Methods

Cluster analysis (CA) comprises a number of methods which differ in (i) the similarity measure for comparing objects: (ii) the mathematical strategy for grouping objects, and (iii) the criterion to establish when a set of objects is a cluster.

Similarity coefficients were grouped by Sneath and Sokal[7] into four main classes, namely distance, association, probabilistic and correlation. The choice of a measure type is strictly dependent on the nature of the data. Willet and Winterman[8] highlighted that, when using fragment bit-strings, correlation and association coefficients give better results than distance. However, when using nonbinary variables like molecular properties, distances are undoubtedly the most commonly used similarity measure, as they allow for a simple geometrical interpretation.

Depending on their mathematical strategy CA methods are divided in two major groups, namely hierarchical and nonhierarchical. Hierarchical agglomerative clustering[2–4] is probably the most widely used method. It starts with n objects in n separate clusters and, after $n-1$ fusion steps, ends with all n objects in one single cluster. At each step the number of clusters is decreased by one, by joining the two closest clusters. A hierarchy of partitions is thus produced, typically represented by a dendrogram. Alternatively, nonhierarchical methods are single-step algorithms as they provide a single optimal partition of data into clusters based on a set of target criteria. These methods can be further classified into density and optimization. Density methods[9–11] take advantage of the intuitive statement that a cluster is observed when the density of objects is locally higher in one place than in other regions. Optimization methods,[2–4] instead, seek a partition of objects into a predetermined number of clusters optimizing a certain criterion (e.g., minimizing the sum of squared distances between each object and its own cluster center).

CA methods have been used extensively in analytical chemistry[2] and, more recently, in the field of molecular modeling[12–23] and molecular diversity.[24–28] Comparative studies between different clustering methods have also been per-

formed.[2,3,23,25,26,28] The reported results highlight that no single method can be claimed superior as its performance is strictly dependent on the nature of the data. For instance, among hierarchical agglomerative methods, single linkage is well known to be affected by a problem called ''chaining''; that is, the inability to distinguish between two clusters connected by a chain of objects.[2,3,23,25] On the other hand, this could represent an advantage when analyzing data sets for which linear clusters occur. In contrast, complete-linkage and several nonhierarchical optimization algorithms work well when data are structured in spherical clusters as they are biased toward this shape even when ''natural'' groupings are characterized by another shape.[2,3] Density clustering methods, as well as Ward's and average-linkage hierarchical algorithms, are less influenced by cluster shape and are therefore said to be able to select ''natural'' clusters.[2]

When it is not known *a priori* whether data are clustered well or not, hierarchical algorithms are preferred to nonhierarchical ones, as the visual analysis of the dendrogram provides a straightforward interpretation of the data tendency toward clustering. On the contrary, nonhierarchical methods may be preferred when data are supposed to be structured. However, several trials are required to obtain the optimal data partition because the setting of input parameters is not trivial.[26] For instance, with the optimization methods the user has to give the number of clusters and this is rarely known *a priori*. The ART-2′ program[12] may overcome this problem, as it optimizes the cluster assignment subject to a constraint on the cluster radius, but the final results may be dependent on the order of input data. The problem of finding the optimal number of clusters has its counterpart in the hierarchical methods in the choice of a threshold distance for ''pruning'' the dendrogram. Several indices have been proposed to make this procedure easier when large amounts of data are analyzed.[22,29] However, their usefulness is questionable when not all clusters are well separated. Furthermore, we agree with Allen et al.[25,26] who pointed out ''optimal clustering'' is essentially a subjective judgment, which needs to be made primarily on the grounds of ''chemical sensibility.'' Therefore, whatever method is used, it is necessary to examine the complete listings of all clusters at various conditions before arriving at a final decision.

SONHICA Clustering

We were interested in using CA to investigate data sets of molecular conformations arising from a conformational search or collected from a dynamics trajectory. In particular, we were aimed at the identification of a technique that would allow us to detect ''natural'' groupings. These could be defined as: (i) those conformations which lie within the same basin on the energy surface, or (ii) those conformation that, although belonging to different energy basins, show a common spatial disposition of several important chemical functionalities. In addition, we were interested in an algorithm that would allow us to take advantage of our ''chemical sensibility'' in a straightforward way. On the basis of these considerations, we developed a new program, SONHICA, with the following main features:

1. The similarity condition that defines whether two or more objects belongs to the same cluster is the following: two objects are considered similar only if all the pairs of descriptor variables differ for less than the chosen thresholds. These thresholds can be assigned individually to the different descriptors,[19] which allows simultaneous handling of a large variety of data descriptors. This procedure is not very different from scaling the input variables, but it is easier and more intuitive, in particular when the input data matrix contains periodic variables (e.g., dihedral angles) or heterogeneous sets of variables (e.g., angles and interatomic distances together).
2. All pairs of objects belonging to the same cluster satisfy the similarity condition. This condition is biased toward finding globular clusters whose dimensions are determined by the assigned thresholds. Based on our experience and literature data,[23,25,26] we believe that the aforementioned cluster definition is suitable for conformational data sets.
3. Clusters are formed going from the most to the least populated, or, more precisely, from the most to the least dense. This criterion can be explained by the following example illustrated in Figure 1a–c. A set of conformations

derived from a dynamic run is shown on a simplified one-dimensional energy surface (Fig. 1a). Because low energy states should be more populated than high energy ones, the conformations are scattered in such a way that the bottom valley is more populated than the walls. Let us consider a threshold value of 60°. By applying criterion 3, objects are grouped as shown in Figure 1b. A large eight-element cluster (dark circles) and two small ones of little significance (white and gray circles) are formed. These data could be grouped in several other ways which would satisfy the same threshold. For instance, in Figure 1c, two clusters (dark and white circles), similarly populated, and one single object (gray circles) are shown. Only by following the highest population criterion, the largest cluster is centered on the potential energy basin and the average conformation is a meaningful representative structure of this conformational state.

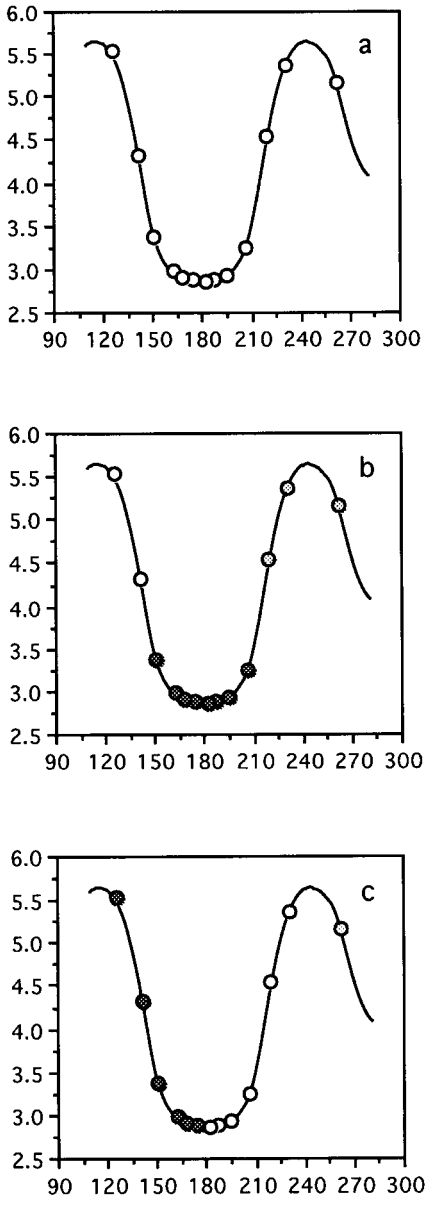

FIGURE 1. (a) Distribution of structures over a simple energy surface. (b) SONHICA clustering for a threshold of 60°. (c) Another possible clustering based on the same threshold value.

INPUT FILE

The input file that SONHICA requires is an ASCII file, whose rows correspond to the objects. The first column represents the identification number (ID) and columns from 2 to $n - 1$ contain the variable values. The last column is optional and represents the ''quality'' associated with the object. Quality is a conformational property, for example, the potential energy of a molecular conformation which can be used to select between equivalent objects, as explained in what follows.

As illustrated in Figure 2, the SONHICA strategy consists of three steps. Let us suppose we have n objects described by p different variables. In step 1, from the input data matrix $D(n \times p)$, the principal matrix $P(n \times n)$ is obtained, which contains the result of all possible comparisons. In step 2,

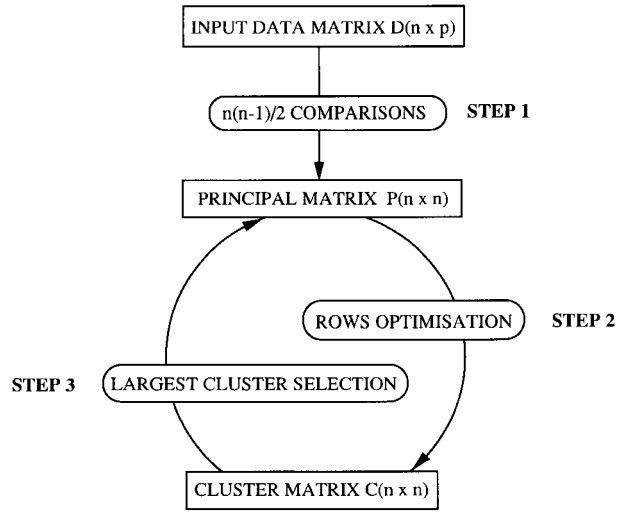

FIGURE 2. Schematic representation of the SONHICA strategy.

matrix P is optimized giving the cluster matrix $C(n \times n)$ containing a first partition in clusters. From the latter the highest populated cluster is selected in step 3. The objects belonging to this cluster are excluded from matrix P, which is then reoptimized. Steps 2 and 3 are repated until all the objects are grouped. The complete procedure is detailed in what follows.

STEP 1

Each object (row of matrix D) is compared to all other objects to give matrix P. Let x_{ij} be the value of variable j for object i, x_{kj} the value of the same variable for object k and tol_j the threshold assigned; objects i and k are defined as similar if:

$$d_j = |x_{ij} - x_{kj}| < tol_j \qquad \text{for each } j = 1 - p$$

when dealing with angles, the following transformations are applied to take into account their periodicity:

$$d_j = |x_{ij} - x_{kj}| \qquad \text{if } |x_{ij} - x_{kj}| <= 180$$
$$d_j = 360 - |x_{ij} - x_{kj}| \qquad \text{if } |x_{ij} - x_{kj}| > 180$$

$n(n-1)/2$ comparisons are made and matrix P is filled: if two objects satisfy the similarity condition a value of 1 is reported, otherwise a 0 is reported. Matrix P is shown in Figure 3, each row representing the temporary cluster centered on the diagonal element. Matrix P is symmetrical and all diagonal elements are obviously unitary. The temporary cluster centered on object no. 1 is illustrated on the right-hand side: object nos. 2, 3, 6, 7, 8, and 9 have results similar to object no. 1, but they are not necessarily similar to each other.

STEP 2

Matrix P is optimized to give matrix C, that is, each row of P is transformed into the most populated cluster. In Figure 3, step 2 is illustrated for the first row (temporary cluster centered on object no. 1) of matrix P. A temporary matrix is created, removing those rows and columns (in gray) corresponding to all the objects not similar to object no. 1. This matrix does not yet satisfy the third criterion mentioned previously as it contains zero elements (i.e., objects not similar to each other); therefore, other objects are still to be removed. This choice is made to find out, in a stepwise manner,

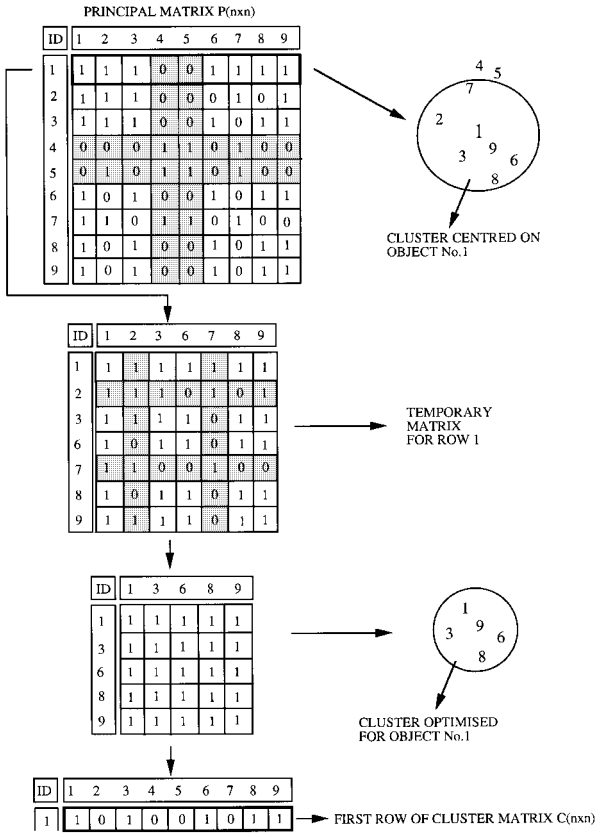

FIGURE 3. The principal matrix, P, which contains the results of all $n(n-1)/2$ comparisons, is optimized, row by row, to give the cluster matrix C (see text for details).

the largest submatrix whose elements are all 1 (i.e., all zero elements must be removed deleting the lowest number of objects). At each step, SONHICA removes the object corresponding to the column that contains the lowest number of unitary elements. When two or more objects are equivalent the choice is made following one of the following two criteria:

- the ''worst quality'' (if quality is the potential energy, the object with the highest energy value is removed); or
- the highest distance from the center of the temporary cluster (if different kinds of descriptors are used, the distance is computed on the group of homogeneous variables preset by the user).

In this way it is possible to assign the highest number of objects to each cluster, independently

from the order of input data. In Figure 3, the temporary cluster centered on object no. 1 is reduced, by first excluding object no. 7 and then object no. 2, to a smaller cluster where all pairs of objects satisfy the similarity condition. All the rows are submitted to the same procedure to obtain matrix *C*.

STEP 3

From matrix *C* the highest populated cluster is then selected. When two or more clusters are equally populated, the user can choose between:

- the cluster containing the object with the ''best quality'' (if quality is the potential energy, best means lowest); or
- the cluster showing the lowest mean standard deviation (if different kinds of descriptors are used, the mean standard deviation is computed on the group of homogeneous variables preset by the user).

Once the most populated cluster is identified, the objects belonging to that cluster are removed from matrix *P* (i.e., the corresponding rows and columns are set to zero) and steps 2 and 3 are repeated on the matrix which is modified until all the objects are assigned. Removing the objects of the selected cluster and repeating the optimization step allows for the identification of the best data partition at each step.

OUTPUT FILE

SONHICA lists all the clusters found ranked from the most to the least populated and for each cluster provides the user with accurate statistics on each group of variables and on the ''quality'' column. These values can be checked by the user to evaluate the significance of each cluster.

Representative objects, that is, the object characterized by the ''best quality'' value and the object closest to the center of the cluster (if different kinds of descriptors are used, the distance is computed on the group of homogeneous variables preset by the user), are also reported.

The output file contains a comparison matrix; that is, all the clusters represented by their centroids or by their centrotypes are compared to each other on the basis of the initial set of threshold values. A symmetrical binary matrix, like *P*, is then produced (see the Results section). The analysis of this matrix allows the user to gain a better idea of the relationships between the different clusters found and could provide useful information about the proper setting of input thresholds.

COMPUTATIONAL DEMANDS

SONHICA is written in the ANSI C programming language within a SGI Unix environment. Like the Lance–Williams implementation[2,3] of hierarchical agglomerative algorithms, memory storage increases as the square of the number of objects. Therefore, when the number of objects is larger than 10^4 it is advisable to use alternative methods which are not based on a $(n \times n)$ matrix. The evaluation of the computational time is more critical because it is strongly influenced by the structure of the data; that is, data sets which tend to form well separated clusters require much less time with respect to data homogeneously spread over the variable space. Also, for this reason, different sets of thresholds lead to different processing times and the minimum time usually indicates the set that produces the most separate clusters.

In all our trials, computed on different data sets, the computational time was always comparable or even smaller than that of the same analysis performed with the hierarchical complete-linkage method.

In general, the program is much faster if the ''quality'' column is used to choose between equivalent objects or clusters. Another way of saving time is to choose a cluster population threshold under which no more clusters are generated. If, for instance, the user is clustering thousands of objects and the highly populated clusters contain hundreds of elements, he/she may not be interested in those clusters populated by less than, say, ten objects. In this case, it is possible to stop the program and exit when the last ten-membered cluster is created.

Finally, the number of descriptors only slightly affect SONHICA, as, once matrix *P* has been filled, float input data are used only in ambiguous situations and to compute the final statistics.

Methods

COMPUTATIONAL METHODS

Model building and conformation analysis were carried out within the Sybyl 6.03 molecular model-

ing package.[30] DGEOM[31] (implemented in Sybyl 6.03) was used when dealing with NMR distance constraints. All the energy calculations were done using the AMBER[32] all-atom force field *in vacuo* with a dielectric constant of $\varepsilon = 40$, and geometry optimizations were carried out by means of Powell's algorithm. CA was performed using SONHICA and the hierarchical complete-linkage algorithm present in Sybyl 6.03 (for comparison). The visual analysis of conformers clustered and properly superimposed was carried out through an interface between Sybyl and SONHICA developed for this purpose and written in Tripos SPL (Sybyl programming language). All calculation were run on a Silicon Graphics Crimson workstation.

CONFORMATIONAL ANALYSIS

(a) ITF1697: H—Gly^1—N(Et)Lys^2—Pro^3—Arg^4—OH. Distance constraints and ^{1}H resonances were assigned from NMR experiments (DQF-COSY, TOCSY, ROESY) performed at 500 MHz on a Bruker AM-500 spectrometer.[5] A total of 200 structures were generated by DGEOM using 39 NMR distance constraints. Molecular conformations were further minimized until the rms gradient was less than 0.01 kcal/mol · Å.

(b) BQ123: *cyclo*(dAsp^1—Pro^2—dVal^3—Leu^4—dTrp^5). Molecular conformations for BQ123 were generated by using a combination of systematic search, molecular dynamics, and energy minimization:

1. Cyclopentapeptide (dAla^1—Pro^2—dAla^3—Ala^4—dAla^5) was investigated by systematically varying ϕ and ψ backbone angles on a 30° grid (the dAla^1 amide bond was taken as ring closure). A total of 143 structures were generated.
2. Conformations were minimized until the rms gradient was less than 0.01 kcal/mol · Å and an rms threshold of 0.25 Å was applied on backbone atoms to avoid tight similar structures. The appropriate side chains were then added to obtain *cyclo*(dAsp^1—Pro^2—dVal^3—Leu^4—dTrp^5) and the Sybyl ''scan'' option was used to avoid bad van der Waals contacts. A total of 34 conformations were obtained.
3. Each conformation was taken as the starting structure for a 100-ps MD at 300 K with a timestep of 1 fs. For each MD run, a structure was stored every 5 ps, giving a total of 714 conformations, which were further slightly minimized.

Results and Discussion

SONHICA clustering was tested on two sets of conformers: (a) 200 structures of ITF1697, a modified linear tetrapeptide, H—Gly^1—N(Et)Lys^2—Pro^3—Arg^4—OH; and (b) 714 structures of BQ123, a cyclic pentapeptide, dAsp^1—Pro^2—dVal^3—Leu^4—dTrp^5.

ITF1697

The conformers generated by DGEOM were clustered on the basis of backbone ϕ and ψ dihedral angles.

The input file for SONHICA contained the ID column (numbers from 1 to 200), six columns corresponding to the values of the dihedral angles ($\psi 1$, $\phi 2$, $\psi 2$, $\phi 3$, $\psi 3$, $\phi 4$, indexed by residue position), and a ''quality'' column, which was represented by the sum of NMR constraint violations. Thus, as the 200 DGEOM structures were derived from an NMR study, the definition of the ''best'' structures, in this case, is strictly linked to their agreement with experimental data. The results obtained were also compared with those from hierarchical complete-linkage clustering.

Three sets of thresholds for the six dihedral angles were tested with SONHICA: (i) all thresholds of 30° each; (ii) all thresholds of 60° each; and (iii) variable thresholds smaller in the center of the molecule and larger at the N- and C-terminal 100–60–30–30–60–100, to take into account the higher flexibility of the terminal groups.

With all thresholds of 30°, SONHICA generated very small clusters, the most populated being a cluster of ten elements, and 67% of the structures were contained in clusters populated by less than five elements.

Raising the thresholds to 60°, the most populated cluster contained 21 molecules, but the quality of the clusters was low. The largest cluster is illustrated in Figure 4a. It is easy to see that a few conformations show a spatial orientation of the CONEt moiety of Lys^2 different from most of the structures. Comparable or even worse results were obtained using the hierarchical complete-linkage algorithm by pruning the dendrogram so that the most populated cluster contained at least 20 structures (Fig. 4b).

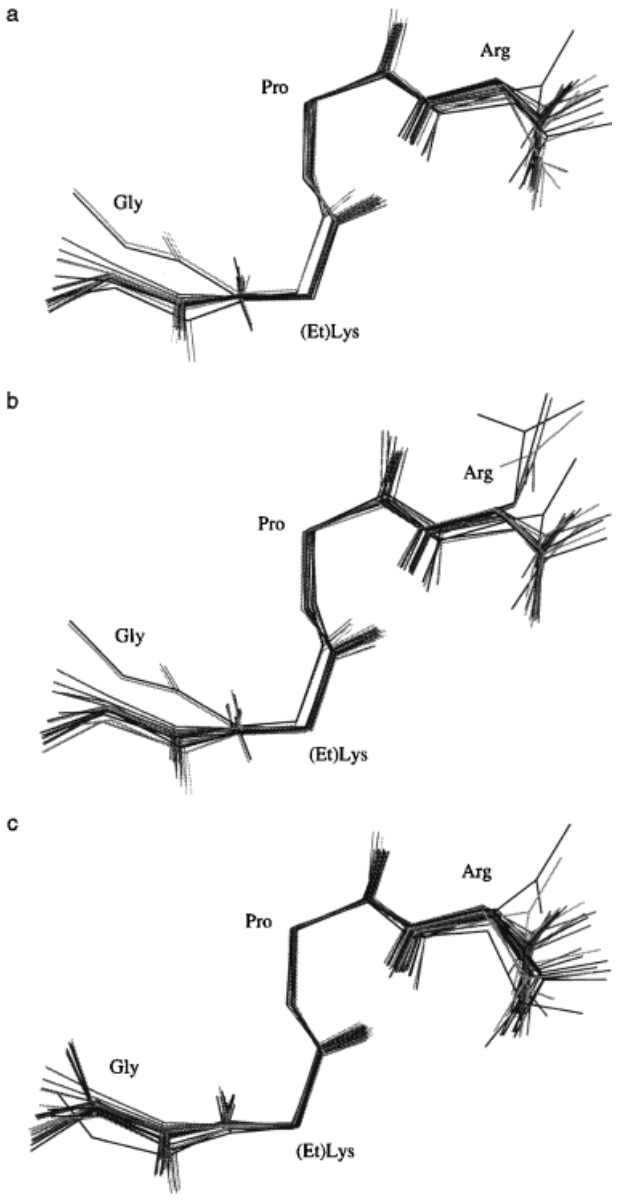

FIGURE 4. Backbone superposition of conformers of ITF1697 belonging to the highest populated cluster obtained with (a) SONHICA (21 structures) by using all 60° tolerances, (b) the complete-linkage hierarchical clustering method (the dendrogram was pruned so that the most populated cluster contained at least 20 structures), and (c) SONHICA (30 structures) by using gradual tolerances (see text for details).

A better result was obtained with SONHICA employing the third set of gradual threshold values. In this case, a cluster of 30 elements was obtained, about 47% of the molecules were included in clusters populated by more than 10 elements, and the quality of the clusters was improved. Thus, all the structures belonging to the largest cluster (Fig. 4c) share the same orientation for the CONEt moiety. This can be evaluated quantitatively by comparing the standard deviation of the ϕ angle of Lys^2 for the clusters shown in Figures 4a and c. Although the threshold on this angle was the same (60°), the standard deviation values were different, being 13.4° (Fig. 4a) and 7.4° (Fig. 4c). This improvement in cluster definition is mainly due to the lower threshold applied to the Lys^2 ψ-angle (30°). In this way, the leverage effects problem was avoided. Thus, as highlighted by Karpen et al.,[12] small changes in the dihedrals near the center of the geometry of an extended structure can cause rather large changes in conformation, whereas this is not true for rotatable bonds positioned at the ends.

Clustering algorithms based on one distance could probably provide equivalent results by applying higher weights to the central dihedral angles than those applied to the terminal angles. However, the periodic nature of input data makes the scaling procedure difficult and there is a risk of producing artifacts. On the contrary, SONHICA multiple thresholds allow the user to give a different weight to each descriptor variable in a more straightforward way. For each of the three SONHICA runs the time required was always less than 6 s.

In some cases, one of the goals of CA is to select a subset of representative structures to perform further investigation. Due to the globular shape of clusters obtained from molecular conformations described by geometrical features, the cluster center (i.e., the average conformation) can be considered a meaningful point to adequately represent all structures belonging to that cluster. The centrotypes (i.e., the closest structures to the vector of the means) are illustrated in Figure 5 for the clusters (populated by at least five elements) created by SONHICA with the gradual thresholds. They are ranked according to their population so that cluster no. 1 contains 30 elements and cluster no. 9 only 7. It is worth noting that a few structures differ in the orientation of the terminals only; however, this result was expected as the ψ- and ϕ-angles of Gly^1 and Arg^4, respectively, were included in the clustering process. The first six structures are characterized by a S shape, where the frame Lys^2—Pro^3 is highly conserved. None of these conformations is characterized by a clear γ-turn around the proline residue. However, hydrogen bonding between the Lys^2CO and Arg^4NH groups stabilizes structures 7, 8, and 9 in a U-shaped, more folded conformation. For all the conforma-

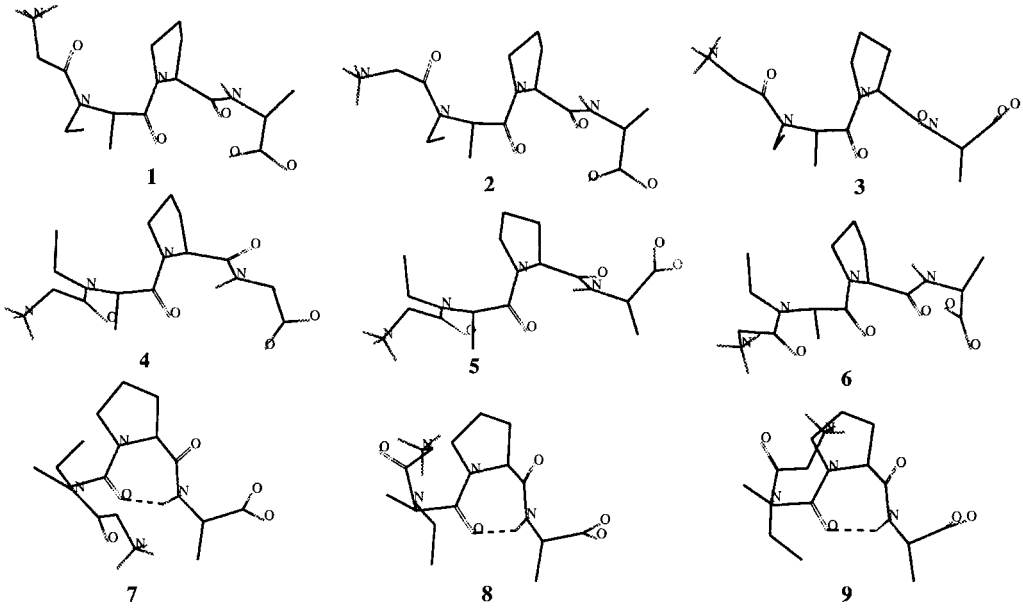

FIGURE 5. Representative ITF1697 structures (centrotypes) of the first nine clusters (populated by at least five elements) created by SONHICA using gradual thresholds.

tion shown, the N-terminal portion, as well as the Et group, explores two well separated regions, whereas the C-terminal region seems to be more variable.

From the analysis of the clusters obtained along with their representative conformations, SONHICA appears able to group similar structures efficiently, as well as to separate significantly different ones, providing useful information for future investigation. Thus, each of these conformations could be used as a starting point for a dynamics run to evaluate their conformational stability or to look for possible interconversions between different families.[5]

BQ123

On 714 conformers of the molecule two different sorts of clustering were applied: the first one using the backbone angles to study the conformational preferences, and the second using the distances between side chains to investigate the spatial disposition that possible 3D pharmacophores could assume.

Backbone

The input file for SONHICA contained the ID column, 11 columns with the ϕ and ψ backbone angles for the five residues and the ω-angle of dAsp[1], and the ''quality'' column with the potential energy of the conformers. The same threshold was applied to the whole set of torsions and five clustering runs with different threshold values, from 40° to 120° with a stepsize of 20°, were executed. The data partition into clusters is shown in Table I; clusters are divided into two groups, depending on the cis or trans configuration of proline, and they are ranked according to the population found by using the 40° threshold (i.e., At and Ac are the most populated clusters out of all the trans and cis proline ones, respectively).

The use of different threshold values allows us to evaluate cluster definitions and to obtain additional information about relationships between different clusters. For example, cluster At (Table I) does not show relevant changes when the threshold value is incremented, suggesting that it could represent a well-separated compact cluster, whereas conformers belonging to cluster Dt join either cluster Bt or Et at 60°; these latter ones merge in a unique cluster at 80°. Similar observations regarding relationships between clusters can also be carried out by looking at the output comparison matrix, represented in Figure 6, computed on trans proline cluster centroids derived from the 40° SONHICA run. Thus, At does not show neighbor clusters, whereas Dt is close to both Bt and Et

TABLE I.
Distribution of BQ123 Conformers into Clusters (Populated by At Least Ten Elements) on the Basis of ϕ and ψ Backbone Angles at Different Threshold (T) Values.[a]

Cluster	T40	T60	T80	T100	T120
At	52	62	66	67	67
Bt	51	69	95	110	113
Ct	21	21	21	21	21
Dt	17	Bt / Et	Bt	Bt	Bt
Et	15	21	Bt	Bt	Bt
Ft	13	18	24	28	29
Gt	11	24	31	36	37
Ht	—	14	18	18	19
It	—	14	16	16	16
Lt	—	—	12	19	21
Ac	55	55	56	58	69
Bc	42	52	54	52	114
Cc	34	44	53	70	Bc
Dc	30	32	32	34	35
Ec	28	33	34	34	39
Fc	23	23	23	23	25
Gc	16	17	16	Cc	Bc
Hc	12	13	Cc	Cc	Bc
Ic	12	13	14	12	Ac
Lc	—	—	—	12	18
No. clusters	16	17	16	16	14
No. objects	432	525	566	610	623

[a] Clusters are divided in two groups, depending on the trans or cis (suffix t or c, respectively) configuration of proline, and they are ranked according to their population found in the first analysis at 40° threshold.

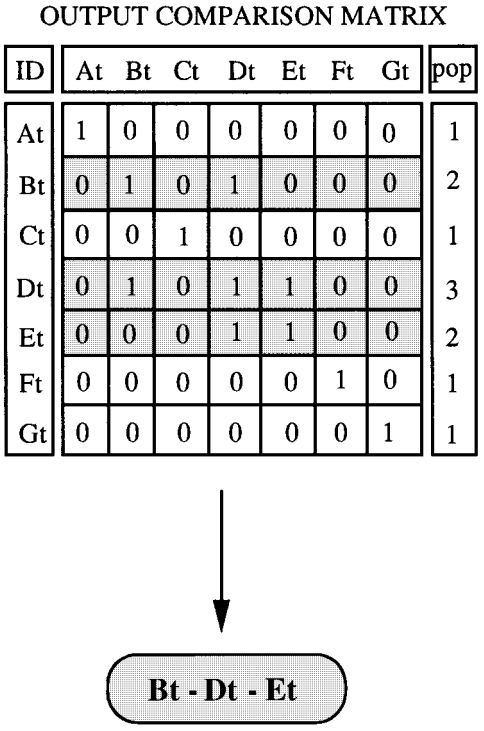

FIGURE 6. Output comparison matrix computed on trans proline cluster centroids derived from the 40° SONHICA analysis of BQ123 backbone angles. Neighborhood relationship are also shown (see text for details).

and seems to represent an intermediate situation, because clusters Bt and Et are not similar to each other. The additional analysis of MD simulation runs indicates that frequent transitions occur between Bt, Dt, and Et, whereas few transitions involve cluster At. On the basis of these observations, we may hypothesize that cluster At corresponds to a sharp, well-separated minimum on the potential energy hypersurface, whereas conformations belonging to Bt, Dt, and Et occupy the same potential energy basin or are separated by a low, thermally accessible barrier.

For each cluster containing at least ten elements at the 40° threshold, the mean angle and energy values are reported in Table II. By looking at energy distribution, clusters At, Bt, and Dt (shown in Fig. 7a, 7b, and 7c , respectively) clearly result to be the most favorable. Thus, these are the only clusters showing an inverse γ-turn around Pro^2, stabilized by hydrogen bonding between the dVal^3 NH and dAsp^1 CO groups. Moreover, cluster Dt is characterized by a ''non ideal'' β turn[33] over residues dVal^3—Leu^4—dTrp^5—dAsp^1, whereas At and Bt present an additional γ-turn centered on dVal^3. As shown in Table II, clusters At and Bt differ by only dTrp^1 ψ and dAsp^1 ϕ dihedral angle values, and Bt contains the global minimum structure out of all the 714 conformers.

The results obtained fully agree with those from the literature. Thus, several NMR studies[34–39] performed in different solvents show that the solution conformation of BQ123 should be characterized by an inverse γ-turn and a loose β-turn like the one found in conformers belonging to cluster Dt. Furthermore, At- and Bt-like structures, which show two γ-turns, were found to be the lowest energy conformers when using the AMBER force field without NMR constraints.[36]

For each of the five SONHICA runs the time required was always less than 60 s.

TABLE II.
Mean Dihedral Angles and Mean and Minimum Energy Kilocalories per Mole for Each Cluster (Populated by At Least 10 Elements) of BQ123 Obtained with SONHICA Using a Threshold of 40° (Population of Each Cluster is Indicated in Square Brackets).

Cluster	dAsp ϕ/ψ	Pro ϕ/ψ	dVal ϕ/ψ	Leu ϕ/ψ	dTrp ϕ/ψ	E_{mean}	E_{min}
At[52]	−56 / −73	−80 / 74	78 / −81	−70 / 98	72 / −122	2.2	0.2
Bt[51]	131 / −86	−79 / 66	80 / −74	−79 / 90	64 / 47	0.9	−1.1
Ct[21]	82 / −122	−66 / −59	−76 / 67	66 / −78	−61 / 91	5.0	4.4
Dt[17]	119 / −91	−73 / 85	75 / −100	−75 / 116	66 / 39	2.0	0.4
Et[15]	85 / −112	−62 / 104	70 / −123	−75 / 109	76 / 77	4.0	2.4
Ft[13]	−52 / −71	−78 / −106	−64 / −81	−78 / 109	77 / −113	5.1	3.8
Gt[11]	136 / −132	−61 / −56	−75 / −122	−87 / 112	70 / 60	5.1	3.9
Ac[52]	70 / 57	−93 / −157	72 / 50	54 / 57	144 / 48	6.7	4.3
Bc[51]	148 / 53	−92 / −164	78 / −86	−68 / −39	15 / 46	4.9	2.1
Cc[21]	62 / 77	−74 / 155	64 / −126	−82 / 64	134 / 31	5.7	3.1
Dc[17]	146 / 57	−88 / −45	−50 / −72	−89 / 64	67 / 45	6.4	4.0
Ec[15]	135 / 74	−79 / 174	72 / 56	75 / −58	−52 / −59	9.2	6.3
Fc[13]	156 / −60	−68 / 152	129 / 58	54 / 57	88 / 105	7.6	5.7
Gc[11]	141 / 55	−89 / −174	77 / −84	−86 / 52	68 / 50	6.5	2.8
Hc[12]	84 / 62	−91 / −175	68 / −100	−83 / 66	82 / 78	8.0	6.4
Ic[12]	139 / 71	−86 / −171	68 / 54	57 / 76	149 / −55	8.6	7.0

Side Chains

From Table I, clusters At and Bt, at the 100° threshold, are populated by 67 and 110 elements, respectively; that is, conformers belonging to Dt and Et have joined Bt. All these 177 conformers were selected and then clustered using distances between side chains.

Structure–activity studies[6] on the ET_A cyclopentapeptide antagonist class showed that residues dAsp[1], dVal[3], Leu[4], and dTrp[5] should directly interact with the receptor active site, whereas Pro[2] should only play a conformational role. However, in this study, all residue side chains were considered as possible pharmacophoric points. Thus, all the ten interatomic distances, which define the relative spatial disposition of the five side chains, each one described by a centroid dummy atom, were used with SONHICA; the potential energy was included as the ''quality'' column.

An additional descriptor variable was introduced in the clustering process: the $\chi 1$ dihedral angle of dAsp[1] (distances and angles can easily be used together with SONHICA), in order to avoid anomalous groupings, such as the one represented in Figure 8; this cluster contains conformers whose dAsp[1] side chains, although close in space, point in opposite directions.

In this second phase, a unique SONHICA run was performed using different threshold values for each descriptor variable. It has been recently highlighted by Greene et al.[40] that the threshold in a geometic constraint of a query for 3D database searches must make chemico-physical sense. It has also been shown that the following variations in position are reasonable for energies likely to be encountered in bound structures.

Hydrogen bonding atoms: 0.5–2.1 Å
$\pi-\pi$ interactions: 0.5–2.0 Å
Hydrophobic interactions: ~ 1.5 Å
Charge interactions: ~ 0.4 Å

For instance, the actual position of a charged center in a ligand could be anywhere within 0.4 Å of the ideal position and still permit a significant interaction. Consequently, the threshold to be applied on a distance measured between two pharmacophoric points must depend on the physicochemical properties of related groups and would be the sum of the positional tolerances shown above.

The threshold values applied to the ten interatomic distances are listed in Table III. In addition, a 60° threshold was applied to the dAsp[1] $\chi 1$ dihe-

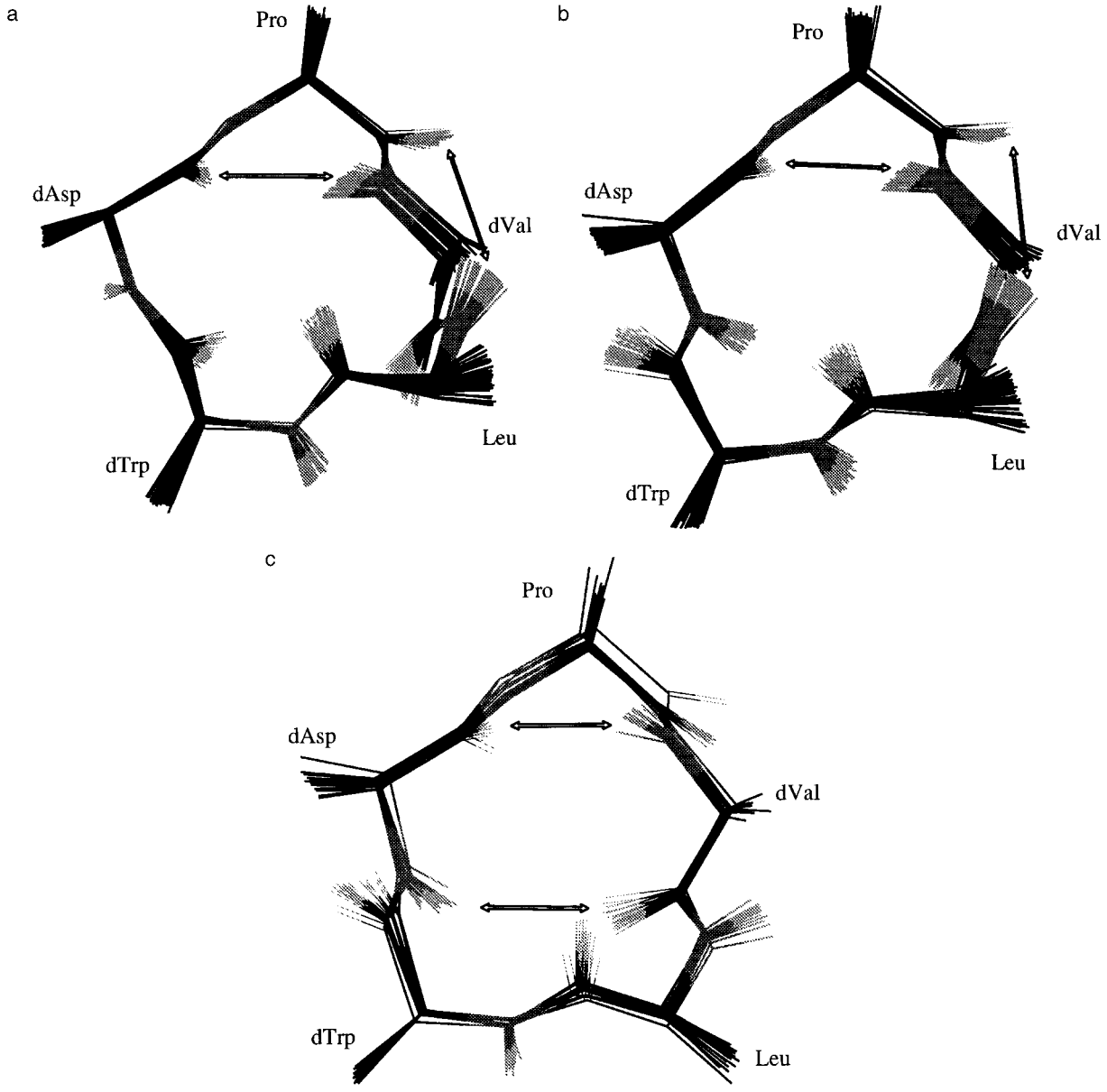

FIGURE 7. Backbone superposition of conformers of BQ123 belonging to (a) cluster At (52 structures), (b) cluster Bt (51), and (c) cluster Dt (17). These clusters were obtained with SONHICA using a 40° tolerance on backbone dihedral angles (see text). Clusters At and Bt are characterized by two γ turns, the first, inverse, around Pro^2, and the second centered on $dVal^3$ (indicated by arrows), while cluster Dt by an inverse γ turn centered on Pro^2 and a loose β turn around $dVal^3$—Leu^4—$dTrp^5$—$dAsp^1$. The latter cluster agrees with the solution conformation of BQ123 found by several NMR studies.[26–30]

dral angle. SONHICA produced ten clusters having more than five elements (time required 5 s), labeled from A to L according to their population. Mean distance and energy values for each cluster are reported in Table IV.

The output comparison matrix computed on cluster centroids, represented in Figure 9, highlights the following similarity relationships: A–F and H–B–C–I. Representative structures (i.e., centrotypes) for the latter clusters are illustrated in Figure 10. By looking in particular at the $dTrp^5$ side chain orientation, cluster B clearly represents an intermediate situation between H and C, whereas cluster C represents a intermediate situation between B and I, with H and I positioned at the ends.

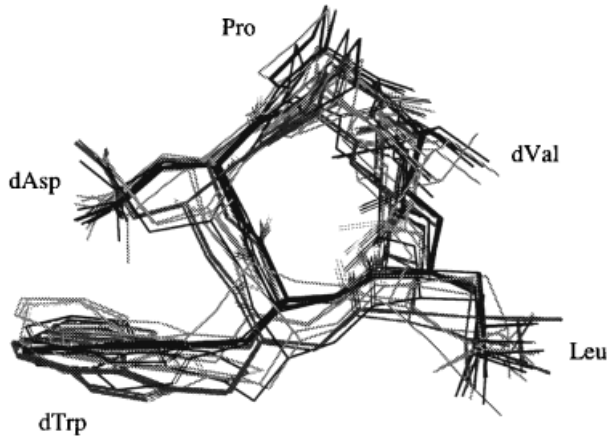

FIGURE 8. Side chain superposition of several conformers of BQ123 belonging to a cluster which was obtained in a preliminary analysis by using the ten interatomic distances without the $\chi 1$ angle of dAsp. We consider this cluster anomalous, because the $dAsp^1$ side chains, although close in the space, point in different directions.

Figure 11 a–c illustrates the three most energetically favored clusters (i.e., B, E, and H). The visual analysis of these clusters highlights that different values of the backbone ϕ and ψ dihedral angles can nevertheless bring the side chains of the molecule into the same regions. These clusters could represent possible 3D pharmacophores of BQ123; however, the comparison with others cyclopentapeptides, showing different affinity values with respect to the ET_A receptor, is necessary to gain more information about the 3D geometrical requirements for the activity.

This kind of application highlights the advantages offered by the SONHICA strategy. Using traditional clustering methods, either nonhierarchical or hierarchical ones, it is necessary to repeat the analysis or the dendrogram prune operation several times, ensuring that the range covered by each descriptor variable within each cluster satisfies the requirements described above. However, with SONHICA, a unique run is requested and analysis of the clusters produced is straightforward and much less time-consuming.

Conclusions

We have described a new nonhierarchical clustering program, SONHICA, which can be successfully employed for analyzing data sets of molecular conformations.

Unlike traditional clustering methods, SONHICA does not make use of an overall index, like a distance, to evaluate similarity between objects, but each descriptor variable is compared individually on the basis of a preset threshold (i.e., two objects are considered similar only if all the pairs of descriptor variables differ for less than the chosen thresholds). This procedure is not very different from scaling the input variables and then using a single distance, but it is certainly simpler and more intuitive. In addition, it provides the user with greater control and sensitivity over the descriptor variables in a more straightforward way.

SONHICA generates clusters with the highest possible density because its iterative procedure creates clusters going from the most to the least dense. It assures high quality clusters due to its cluster definition (i.e., all the objects belonging to a cluster must be similar to each other) and is particularly suitable for data sets which tend to form globular clusters, such as those consisting of conformational features of flexible molecules.

SONHICA has been applied to investigate the conformational space of a modified linear tetrapeptide, ITF1697, and of a cyclic pentapeptide, BQ123. In these applications, the clustering process has been carried out employing dihedral angles

TABLE III.
Threshold Values (Angstroms) Applied to the Ten Interatomic Distances Used to Clusterize 177 Conformers of BQ123 (See Text).

Distance		Tolerance	Distance		Tolerance
d1:	$dAsp^1$—Pro^2	2.0	d6:	Pro^2—Leu^4	3.0
d2:	$dAsp^1$—$dVal^3$	2.0	d7:	Pro^2—$dTrp^5$	3.5
d3:	$dAsp^1$—Leu^4	2.0	d8:	$dVal^3$—Leu^4	3.0
d4:	$dAsp^1$—$dTrp^5$	2.5	d9:	$dVal^3$—$dTrp^5$	3.5
d5:	Pro^2—$dVal^3$	3.0	d10:	Leu^4—$dTrp^5$	3.5

TABLE IV.
Mean Distances (Angstroms), Mean and Minimum Energy (Kilocalories per Mole) for Each Cluster (Populated by At Least Five Elements) of BQ123 Obtained with SONHICA (Thresholds Applied Are Listed in Table III) (Population of Each Cluster is Indicated Square Brackets).

Cluster	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	E_{mean}	E_{min}
A[41]	6.7	9.0	10.6	3.5	7.1	8.2	8.4	7.8	10.6	10.2	1.8	0.2
B[27]	5.6	8.0	11.1	5.0	6.9	7.6	8.4	7.4	8.1	10.5	0.8	−0.8
C[19]	5.2	8.8	11.4	5.6	7.1	9.4	8.5	7.1	11.1	10.5	3.5	1.9
D[15]	5.5	8.6	11.4	10.6	7.2	8.7	10.2	8.3	9.7	4.9	2.7	1.6
E[13]]	6.7	8.6	10.6	6.8	6.8	7.5	8.9	7.7	5.6	8.1	0.7	−1.1
F[11]	6.4	10.2	10.4	3.9	7.0	10.1	9.1	7.3	11.9	10.0	3.8	2.2
G[10]	6.6	9.5	9.9	8.3	7.2	9.1	10.1	8.2	10.6	4.6	2.4	1.4
H[9]	5.5	8.2	11.0	7.5	7.0	7.5	8.4	7.5	5.1	7.9	0.3	−1.1
I[9]	5.6	8.4	11.2	8.0	7.1	8.0	9.2	7.8	11.9	9.6	3.6	2.3
L[6]	6.7	9.3	10.4	8.8	7.1	6.0	9.9	7.7	9.9	7.6	5.4	5.0

OUTPUT COMPARISON MATRIX

ID	A	B	C	D	E	F	G	H	I	L	pop
A	1	0	0	0	0	1	0	0	0	0	2
B	0	1	1	0	0	0	0	1	0	0	3
C	0	1	1	0	0	0	0	0	1	0	3
D	0	0	0	1	0	0	0	0	0	0	1
E	0	0	0	0	1	0	0	0	0	0	1
F	1	0	0	0	0	1	0	0	0	0	2
G	0	0	0	0	0	0	1	0	0	0	1
H	0	1	0	0	0	0	0	1	0	0	2
I	0	0	1	0	0	0	0	0	1	0	2
L	0	0	0	0	0	0	0	0	0	1	1

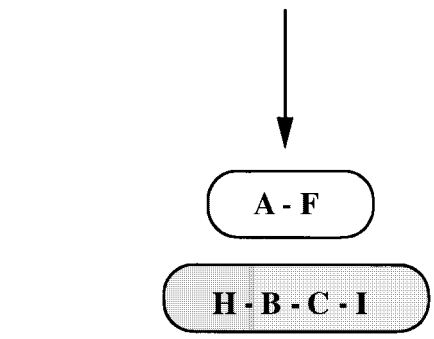

FIGURE 9. Output comparison matrix computed on cluster centroids derived from the SONHICA analysis of BQ123 side chain distances. Neighborhood relationships are also shown (see text for details).

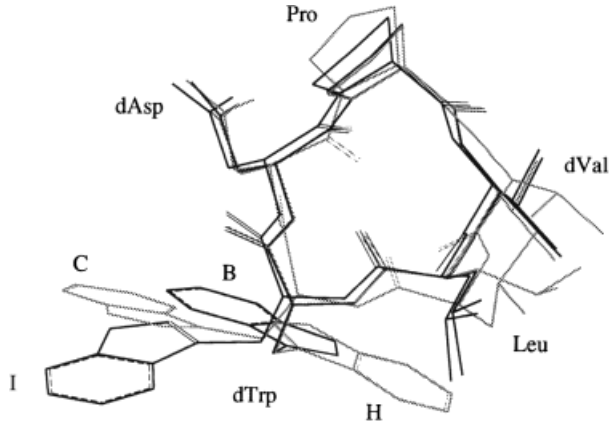

FIGURE 10. Backbone superposition of the representative conformations (i.e., centrotypes) of clusters B, C, H, and 1, which result in being close to each other on the basis of the output comparison matrix of Figure 7.

and interatomic distances. A common disadvantage, when using dihedral angles, is that they differently affect the overall shape of a linear molecule, depending on their distance from the center of geometry. This problem can be overcome by SONHICA, as pointed out by the example on ITF1697, by applying gradual thresholds which decrease when going from the ends to the center of the molecule. A second well-known disadvantage is that drastic changes in successive dihedral angles may have little effect on the overall molecular structure. This is the case, for instance, of a subset of BQ123 conformations which, although sharing a high degree of conformational similarity, have been

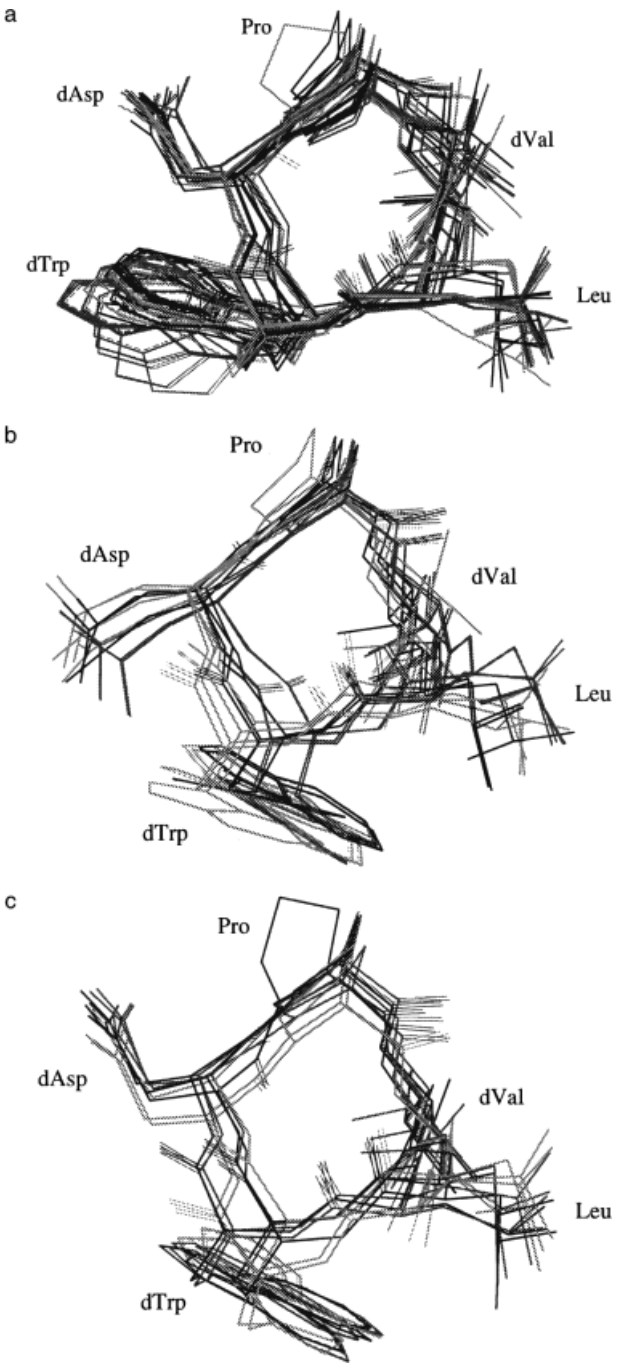

FIGURE 11. Side chain superposition of conformers of BQ123, clustered on the basis of interatomic distances (see text), belonging to (a) cluster B (27 structures), (b) cluster E (13), and (c) cluster H (9).

assigned to different clusters (i.e., At and Bt) (see Table I), because they are characterized by a different specular orientation of the $dAsp^1$ amide linkage. However, the user can easily detect this type of difference between clusters by looking at the statistics included in the output SONHICA file (see Table II).

Interatomic distances are widely used to evaluate conformational similarity, in particular, when the goal of the analysis is to search for possible common spatial dispositions of several important chemical functionalities. For instance, it is well-known that peptide molecules can accomodate side chains in closely related regions, even though they show different backbone patterns. The clustering run on BQ123 using interatomic distances between side chains confirmed this behavior. This application also highlighted the advantages arising from the combined use of heterogeneous descriptors. Conformations, characterized by $dAsp^1$ side chains pointing in different directions, were grouped together (see Fig. 8) on the basis of interatomic distances alone. This is because interatomic distances are slightly influenced by the rotation around the $\chi 1$ $dAsp^1$ side chain. The use of one additional descriptor helped us to avoid this problem: these conformations were assigned to different clusters once this dihedral angle was added.

The combination of angles and interatomic distances could be also useful to overcome the clear disadvantage that interatomic distances are not sensitive to chirality. Otherwise, when expecting chirality problems, the user can employ an RMSD matrix computed after rigid-body superimposition of all the structures on a common frame. Each row of this matrix corresponds to a different structure described by its RMSD values with respect to all the other structures. Two conformations with low RMSD should have similar RMSD vectors. Actually, this kind of matrix can be used as input for SONHICA. For instance, the 177 BQ123 conformations were also clustered using an RMSD matrix, achieving results equivalent to those obtained with interatomic distances (data not shown). It is worth noting that, in this case also, the $\chi 1$ $dAsp^1$ side chain was necessarily added to avoid anomalous groupings (Fig. 8).

There may be two objections to SONHICA as it requires multiple thresholds: (i) it can be difficult to set so many parameters *a priori*; and (ii) when setting a lot of parameters, the output clusters may be function of user preference instead of the real data distribution. We believe that the proper setting of input thresholds is not a difficult problem when analyzing conformational data sets, because a great deal of information is available about the behavior of common geometric descriptors, such as dihedral angles or interatomic distances. For example, thresholds on distances between hypo-

thetical pharmacophoric points can be assigned taking into account the physico-chemical properties of the related groups. Likewise, each torsion shows a characteristic behavior depending on the sequence of the four bonded atoms types. The arbitrary assignment of thresholds is restricted because they can be set on the basis of the functional shape of each descriptor. In addition, the reliability of the data partition can be checked by looking at the statistical parameters computed on each cluster and printed in the SONHICA output file. We believe that these values are useful indicators of the quality of the assigned thresholds.

We are aware that other methods, such as hierarchical algorithms may be preferred when the analysis is exploratory, as no information about data clustering tendency is available. Moreover, when data cluster well, but cluster shape is not globular SONHICA is not advisable and methods less sensitive to the shape have to be used (i.e., density methods). Nevertheless, on the basis of the results presented here, we are confident that SONHICA represents a new and interesting tool in the field of the clustering methods applied to the analysis of molecular conformations.

Acknowledgments

The authors thank Dr. James H. Wikel and Dr. Richard E. Higgs Jr. at Ely Lilly and Co. for useful suggestions in revising the paper. We are also grateful to the unknown referees for their helpful criticism. Dr. Daniela Salvatore and Dr. Claire Thomas are acknowledged for stylistic revision. Finally, we wish to thank Dr. Gianluca Fossati for critical reading of the manuscript.

References

1. A. R. Leach, In *Reviews in Computational Chemistry*, *Vol. II*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, 1991, p. 1.
2. D. L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, Wiley, New York, 1983, and references therein.
3. B. S. Everitt, *Cluster Analysis*, Edward Arnold, London, 1993, and references therein.
4. I. E. Frank and R. Todeschini, In: *The Data Analysis Handbook*, Elsevier, Amsterdam, 1994, p. 38.
5. M. Pegna, A. Biffi, G. Bravi, S. Cappelletti, E. Gancia, M. Pinori, A. Zaliani, L. Zetta, and P. Mascagni, manuscript in preparation.
6. T. Fukami, T. Nagase, K. Fujita, T. Hayama, K. Niiyama, T. Mase, S. Nakajima, T. Fukuroda, T. Saeki, M. Nishikibe, M. Ihara, M. Yano, and K. Ishikawa, *J. Med. Chem.*, **38**, 4309 (1995).
7. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman and Co., San Francisco, 1973.
8. P. Willet and V. A. Winterman, *Quant. Struct. Activ. Relat.*, **5**, 18 (1986).
9. D. Wishart, *Clustan User's Guide*, *The Clustan Project*, University College, London, 1975.
10. D. Coomans and D. L. Massart, *Anal. Chim. Acta*, **133**, 225 (1981).
11. R. A. Jarvis and E. A. Patrick, *IEE Trans. Comput.*, **C-22**, 1025 (1973).
12. M. E. Karpen, D. J. Tobias, and C. L. Brooks III, *Biochem.*, **32**, 412 (1993).
13. H. I. Gordon and R. L. Somorajai, *Proteins*, **14**, 249 (1992).
14. P. Murray-Rust and J. Raftery, *J. Mol. Graphics*, **3**, 50 (1985).
15. R. Unger, D. Harel, S. Wherland, and J. L. Sussman, *Proteins*, **5**, 355 (1989).
16. T. D. J. Perkins and D. J. Barlow, *J. Mol. Graphics*, **8**, 156 (1990).
17. M. J. Rooman, J. Rodriguez, and S. J. Wodak, *J. Mol. Biol.*, **213**, 327 (1990).
18. D. R. McKelvey, C. L. Brooks, and M. Mokotoff, *J. Prot. Chem.*, **10**, 265 (1991).
19. A. Polinsky, M. Goodman, and K. A. Williams, *Biopolymers*, **32**, 399 (1992).
20. D. Gautheret, F. Major, and R. Cedergren, *J. Mol. Biol.*, **229**, 1049 (1993).
21. L. Holm and C. Sander, *J. Mol. Biol.*, **233**, 123 (1993).
22. P. S. Shenkin and D. Q. McDonald, *J. Comput. Chem.*, **15**, 899 (1994).
23. A. E. Torda and W. F. van Gunsteren, *J. Comput. Chem.*, **15**, 1331 (1994).
24. P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Leschworth, UK, 1987.
25. F. H. Allen, M. J. Doyle, and R. Taylor, *Acta Cryst.*, **B47**, 29 (1991).
26. F. H. Allen, M. J. Doyle, and R. Taylor, *Acta Cryst.*, **B47**, 41 (1991).
27. J. M. Barnard and G. M. Downs, *J. Chem. Inform. Comput. Sci.*, **32**, 644 (1992).
28. G. M. Downs and P. Willett, *J. Chem. Inform. Comput. Sci.*, **34**, 1094 (1994).
29. G. W. Milligan and M. Cooper, *Psychometrika*, **50**, 159 (1985).
30. Tripos Associates, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.
31. J. M. Blaney, G. M. Crippen, A. Dearing, and J. S. Dixon, *QCPE Bull.*, **10**, 37 (1990).
32. S. J. Weiner and P. A. Kollman, *J. Comput. Chem.*, **7**, 230 (1986).
33. P. Y. Chou and G. D. Fasman, *J. Mol. Biol.*, **115**, 135 (1977).

34. R. A. Atkinson and J. T. Pelton, *FEBS Lett*., **296**, 1 (1992).
35. S. R. Krystek Jr., D. A. Bassolino, R. E. Bruccoleri, J. T. Hunt, M. A. Porubcan, C. F. Wandler, and N. H. Andersen, *FEBS Lett*., **299**, 255 (1992).
36. E. K. Bradley, C. S. Ng, J. S. Reyna, and D. C. Spellmeyer, *Bioorg. Med. Chem*., **2**, 279 (1994).
37. J. W. Bean, C. E. Peishoff, and K. D. Kopple, *Int. J. Pept. Prot. Res*., **44**, 223 (1994).
38. P. Verheyden, I. Van Asche, M.-H. Brichard, T. Demaude, I. Paye, A. Scarso, and G. Van Binst, *FEBS Lett*., **344**, 55 (1994).
39. M. C. Gonnella, X. Zang, Y. Jin, O. Prakash, C. G. Paris, I. Kolossvary, W. C. Guida, R. S. Bohacek, I. Vlattas, and T. Sytwu, *Int. J. Pept. Prot. Res*., **43**, 454 (1994).
40. J. Greene, S. Kahn, H. Savoj, P. Sprague, and S. Teig, *J. Chem. Inf. Comput. Sci*., **34**, 1297 (1994).